DOCUMENT RESUME

ED 468 522 EA 031 898

AUTHOR	Scudder, David F.
TITLE	An Evaluation of the Federal Class-Size Reduction Program in Wake County, North Carolina--1999-2000.
PUB DATE	2002-04-00
NOTE	19p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; *Class Size; Comparative Analysis; Elementary Secondary Education; Program Evaluation
IDENTIFIERS	*Class Size Reduction; *North Carolina

ABSTRACT

An empirical evaluation of the federal class-size reduction (CSR) program in Wake County, North Carolina, during the 1999-2000 school year is presented. The qualitative process evaluation showed implementation issues involving the mechanics and the meaning of CSR. Often, schools did not understand where CSR occurred because of changing enrollment across grade levels between hiring new teachers and the arrival of the students. Less than optimal results occurred at a few schools where implementation models intended for use with limited space were used to introduce pullout models and other targeted services for at-risk students despite the intent of CSR to move away from these practices. The quantitative evaluation of achievement outcomes used a nonequivalent comparison-group design with analysis of covariance to assess the effects of reduced class size on academic growth in language skills. CSR students in the first and second grades grew more than the comparison group, in some instances equaling the results found in the Project STAR research. Effect sizes were calculated and explained in terms of months of school in larger classes. The analysis showed interaction between CSR and pretest scores, which affected only students with free or reduced-price lunches. Disadvantaged CSR students with low pretest scores performed no better than did comparison group students, but they did perform slightly better than disadvantaged students with average or high pretest scores relative to their comparison groups. (Author/RT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 468 522

AN EVALUATION OF THE FEDERAL CLASS-SIZE REDUCTION PROGRAM IN WAKE COUNTY, NORTH CAROLINA—1999-2000

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

David F. Scudder

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

David F. Scudder, Ph.D.
Social Research and Assessment, Inc.
Data Driven Solutions for Organizations
dscudder@cfl.rr.com
407-963-5319

EA 031 898

Presented at the annual meetings of the
American Educational Research Association (AERA)
April 2002

Note: The author was a Senior Administrator at the Wake County Public School System during the period in which the research presented here was conducted. He is now an independent research contractor operating out of Winter Springs, Florida.

BEST COPY AVAILABLE

ABSTRACT

An empirical evaluation of the federal class-size reduction (CSR) program in Wake County, North Carolina during the 1999-2000 school year is presented. The qualitative process evaluation showed implementation issues involving the mechanics and the meaning of CSR. Often schools did not understand where CSR occurred because of changing enrollment across grade levels between hiring new teachers and the arrival of the students. Less than optimal results occurred at a few schools where implementation models intended for use with limited space were used to introduce pullout models and other targeted services for at-risk students despite the intent of CSR to move away from these practices. The quantitative evaluation of achievement outcomes used a non-equivalent comparison group design with the analysis of covariance to assess the effects of reduced class size on academic growth in language skills. CSR students in the first and second grade grew more than the comparison group, in some instances equaling the results found in the Project STAR research. Effect sizes were calculated and explained in terms of months of school in larger classes. The analysis showed interaction between CSR and pretest scores, which affected only students with free or reduced-price lunches: Disadvantaged CSR students with low pretest scores did not perform as much better than similar comparison group students as did those disadvantaged students with average or high pretest scores. This result obtained despite the finding that growth for disadvantaged students over the comparison group equaled or exceeded that of other types of students.

AN EVALUATION OF THE FEDERAL CLASS-SIZE REDUCTION PROGRAM IN WAKE COUNTY, NORTH CAROLINA—1999-2000.

Introduction

Small classes have always had an intuitive appeal for parents. To many, it appears obvious that smaller classes should be associated with greater achievement. Still, all other things remaining equal, small classes are more expensive than large classes. Recent research suggests that smaller classes actually do improve academic achievement. This paper examines this and other important issues as they relate to the first year (1999-2000) of the federal **Class-Size Reduction (CSR)** program in the Wake County Public School System (WCPSS).

Background

The U.S. Congress authorized the CSR program in 1999 under section 310 of Public Law 106-113. It is the most recent development of the unprecedented interest over the last 15 years in school reform to improve the quality of the nation's public schools. The purpose of the CSR program was to put 100,000 new and fully qualified teachers into America's public schools in order to reduce class size to a national average of no more than 18 in grades one through three. The CSR program is based on a body of high quality experimental research, including Tennessee's Project STAR, which demonstrated that substantial reductions in class size have a significant effect on improving achievement.

For 1999-2000 the U.S. Congress allocated $1.2 billion for the CSR program, enough for about an initial 30,000 teaching positions nationwide. North Carolina received approximately $24.7 million for the 1999-2000 school year. School district allocations were based on the number of children in poverty (80 percent) and total enrollment (20 percent). The allocation for WCPSS was approximately $1.1 million for the 1999-2000 school year.

Wake County Implementation

The objective of the WCPSS implementation plan was to reduce class sizes within targeted schools. In order to accomplish this overall objective, several specific activities were required:

- Hire the maximum number of fully qualified teachers possible with the available funds,
- Determine which schools would receive additional teachers,
- Establish implementation models for deploying the new teachers from among which participating schools could choose,
- Determine which grade levels to target.

Evaluation Questions

Four general questions were addressed in the evaluation:

1) Was the program implemented as planned and, if not, why?
2) What actual services were provided?
3) What were the effects of the program?
4) How could the program be improved?

This paper focuses on the effects of the program on academic achievement.

Implementation Plan

District staff determined that 23 teachers could be supported by the CSR funds. The 23 schools with the most need in terms of three indicators were invited to participate:

- Percent of students receiving free or reduced-price lunches (FRL),
- Number of students (grade 3-8) whose academic achievement was below grade level,
- Percent of students (grade 3-8) whose academic achievement was below grade level.

These 23 schools had between 21.6 and 51.1 percent of their students receiving free or reduced-price lunches (FRL). They also had between 50 and 117 low-achieving students (grade 3-8), which represented between 25.8 and 43.9 percent of the students.

District staff developed four implementation models that reflected the national guidance document published by the U.S. Department of Education. Models 1 and 2 involved adding an additional classroom, while Models 3 and 4 involved having an additional teacher rotate among existing classes to team with the regular teachers at a grade level (see Table 1). District staff recommended the selection of Model 1 unless adequate space was not available for an additional classroom. Schools were asked to implement CSR in grades 1 or 2 (national guidance allowed grades 1-3 except in special circumstances, where grades 4-8 were allowed).

Actual Implementation

All 23 of the invited schools elected to participate. Twenty-three licensed teachers were hired. Students were served in different target grades and by several implementation models. Model 1 (the preferred model) and the second grade were selected by the most schools (see Table 1).

Table 1 illustrates that five schools, usually unintentionally, reduced class sizes in kindergarten or the third grade, which were not within the WCPSS guidelines (grade 3) or the federal statute (kindergarten) for the 1999-2000 year.

The key implementation issue was that many schools did not know which grade level experienced class size reduction. The result turned the placement of CSR into

something of a shell game for schools. The reason was that class size is a moving target that changes throughout the year. A typical example involved a school that hired an extra second grade teacher in July based on its planning data. The plan was to reduce class size in the second grade from around 23 to about 18. By the 10th day of school there were fewer second grade students than expected but more kindergarten students than expected. To compensate for the moving enrollment target, the school moved a second grade teacher (not the one hired with CSR funds) to kindergarten. The result of this change was that the average number of students in the second grade classes became equal to or slightly above the state maximum while the average number of students in kindergarten classes was below the state maximum. Clearly, the effect of class size reduction due to the extra teacher in the school was experienced in kindergarten, not in the second grade where the teacher hired with CSR funds actually worked. But to the school, and reasonably so, CSR occurred where the CSR teacher worked and not where an abstract mathematical reduction in class size occurred from having an extra position.

An even more complicated example, involved the use of mixed-grade classes. The school that implemented CSR in the third grade managed to do this as a result of creating a mixed second and third grade class to compensate for unexpected enrollments, which spread the effect of CSR across two grade levels and included the third grade. The complications arising from the moving enrollment target resulted in weeks of extra qualitative research to determine exactly where CSR was actually implemented. In one instance, a school reported CSR in the first grade (where the CSR teacher actually worked) with average class sizes equal to the state maximum, but at the same time had kindergarten classes well below the state maximum.

Table 1: Number of schools implementing CSR in each model and grade level.

Implementation Model	Kindergarten	Grade 1	Grade 1-2 Combined	Grade 2	Grade 2-3 Combined	Total
1. Teacher of new class **about equal in size to** all other classes of the target grade	4	2	0	8	0	14
2. Teacher of new class **substantially smaller than** other classes of the target grade	0	3	1	1	1	6
3. Rotating teacher shared equally **among all** of the classes of the target grade.	0	0	0	3	0	3
4. Rotating teacher shared equally **among some** of the classes of the target grade	0	0	0	0	0	0
Total	4	5	1	12	1	23

A further implementation issue was that the three schools that selected model 3 did not use the team teaching approach recommended in the guidance document that was provided to each school. Instead, at-risk students were "pulled out" of class to receive tailored instruction (which was not the intention of the federal or local guidelines). The principals at these schools firmly believed they were doing what was in the best interest of the students. District staff informed them that this was not an allowable use of CSR

BEST COPY AVAILABLE

funds and that the research literature did not support this practice. Only one of the schools continued this approach in the 2000-2001 school year, and was, therefore, not awarded a CSR teacher in 2001-2002. Another implementation issue involved the use of Model 2. A few of the schools that adopted Model 2 used the substantially smaller class to target at-risk students even though additional space was available. Although this was not prohibited, it was not expected. It was expected that students in the substantially smaller classes in Model 2 would have approximately the same heterogeneity as regular classes. Class heterogeneity was required for the 2000-2001 school year. These results suggest that many school administrators are wedded to the "squeaky wheel gets the grease" model in which extra resources are applied to those units that need the most improvement, whereas the entire CSR literature implies a different model, in which the regular flow of ordinary social activities in smaller classes is most beneficial to those children that need improvement most.

The 23 teachers hired under the CSR program enabled reduced size classes to be offered to 2,473 students as of the 10th day of the school year; about 107 students per teacher hired. As depicted in Table 2, the number of students served in each implementation model and grade level mirrored the number of students in the targeted grade levels.

Table 2: Number of students served for each implementation model and grade level.

Implementation Model	Kindergarten	Grade 1	Grade 2	Grade 3	Total
1. Teacher of new class **about equal in size to** all other classes of the target grade	485	166	840	0	1491
2. Teacher of new class **substantially smaller than** other classes of the target grade	0	339	227	121	687
3. Rotating teacher shared equally **among all** of the classes of the target grade.	0	0	295	0	295
4. Rotating teacher shared equally **among some** of the classes of the target grade	0	0	0	0	0
Total	485	505	1362	121	2473

As depicted in Table 3, the amount of CSR that was achieved varied by implementation model, with the most reduction achieved under model 1 and the least under model 3. This result obtained because it is only under model 1 that all students in a grade level receive the maximum and equal benefits from CSR all day, every day. All other models mean that the average student receives less CSR benefits than they would have under model 1.

Table 3: Class-size reduction achieved for each implementation model.

Model	Students Served	Average Before	Average After	Average Reduced
1	1491	24.05	**19.62**	4.45
2	687	26.08	**21.46**	4.59
3	295	24.58	**22.42**	2.17

BEST COPY AVAILABLE

Adding one teacher to each grade level did not result in the achievement of classes of the size recommended by the experimental research literature (12-15) or the enabling legislation (18), even in model 1, because the average class size before adding the additional teacher exceeded 23. Twenty-three students is the minimum allowed by the state allocation formula, not the maximum in grades K-2. The maximum is 26 students per class in grades K-2. Most classes are above the minimum. In order to reduce the average class size for each student to 18 in grades K-3, at least one and often two teaching positions would have to be added per grade level using Model 1, depending on the size of the school. Careful attention to the total number of students in each grade in each school would be required to keep classes from drifting well above the target of 18.

While most of the 23 participating schools had enough space to create one additional class, they would not have had the space to create one or even two additional classes for each grade level; at least not without re-designing the existing spaces for more classrooms with fewer students.

Methodology

The evaluation of achievement effects used a pre- and posttest, non-equivalent comparison group design. Because students were not randomly assigned to small and regular classes, the comparison group was not precisely equivalent.

Comparison Group

A comparison group, composed of similar students that did not receive program services must be identified in order to assess the effects of a program such as CSR. When comparison groups are properly constructed it is possible to attribute to the program observed differences between the comparison group and the group that received the program services.

Random sampling was used to construct a comparison group equal in size and demographic composition to the students who received CSR. Separate, equal size, simple random samples of students not receiving the program were drawn corresponding to each demographic segment of the students that received CSR (e.g., white girls with reduced-price lunches or African American girls with full-price lunches) for each implementation model and grade level. The traits used to draw the sample of comparison group students included grade level, implementation model, gender, ethnicity, and whether or not a student received a free or reduced-price lunch. Thus, the students who received CSR under implementation Model 1 in the first grade were compared to other students in the first grade that had the same demographic characteristics.

The schools that received CSR were those with the most disadvantaged students. It would not have been appropriate to construct comparison groups from the district as a whole, even though the students selected had the same demographic characteristics. Students that do not receive free or reduced-price lunches (FRL) in other schools tend to come from families with higher socioeconomic standing than do those in the 23 target schools that do not receive free or reduced-price lunches (FRL). Similarly, those

receiving free or reduced-price lunches (FRL) in the 23 target schools may be more disadvantaged than those receiving free or reduced-price lunches (FRL) elsewhere.

Disadvantaged students in grades 3-8 from schools with relatively low proportions of disadvantaged students tend to have greater academic growth on end-of-grade tests than do similar students from schools with relatively high proportions of disadvantaged students. The 23 selected schools tend to have higher proportions of disadvantaged students than do the other schools. These are systematic and non-random school (neighborhood) level effects that could bias the equivalence of the students for the purpose of comparison by affecting the expected amount of growth without any additional services.

In order to control for school-level effects, the comparison students were drawn as much as possible from the same 23 schools that received CSR. That is, the pool of potential comparison group students was limited to the same 23 schools that received the CSR teachers. Schools that implemented CSR in the first grade, for example, provided comparison group students for those schools that implemented CSR in the second grade.

Since there were several measures of achievement outcomes, which could not be joined together well in a composite score, no pretest scores were used to construct the matched comparison group. The effort to sample the comparison group students from the same schools as the CSR students was, in part, intended to help ensure that there would be no significant differences in pretest scores for any measure of achievement between the CSR and the comparison students. The size of the classes before adding the additional teaching position could not be used to match the comparison students with the CSR students. This occurred because the size of a student's class could not be accurately calculated with the computer system available in the district. The size of the CSR classes both before and after the addition of the new teacher was obtained by a special survey of the participating schools.

Two of the 23 target schools had to be removed from consideration for drawing comparison group students because they were the recipients of another grant that lowered their class sizes in all grade levels substantially below the county average, even below the level most other schools were able to achieve with the CSR funds. These schools received a CRS teacher because they remained among the 23 eligible schools. The technique was successful for all but a few small demographic strata (mostly FRL students), which required expanding the pool of students available to construct the comparison groups beyond the 23 eligible schools in order to find enough FRL students to construct the comparison groups for those strata. The number of comparison students selected from outside these target schools was very small.

Because the pool used to construct the comparison group had to be expanded beyond the 23 target schools, some anomalous findings became quite perplexing. It remains possible that the anomalous findings arose from the departure from the plan for generating the comparison groups. This issue is fully addressed below.

Obviously, non-equivalent comparison groups constructed in this manner cannot control for all factors that may influence outcomes. Only fully experimental randomized designs

can effectively control for all unmeasured traits. However, comparison groups constructed in this manner represent sound quasi-experimental controls for threats to internal validity.

Missing information for pre or posttest scores make the actual number of cases lower than the number reported as served. For example, as of the 10th day of school, 166 students were shown to be receiving CSR services in first grade classes. At the end of the school year, only 135 CSR and 132 comparison students had both pre and post test data. Some of these students left the district in the middle of the year, while others had not been in the district in the previous year. For a few no scores were submitted.

Analysis

A between subjects, pre- and posttest design was used to test the effects of the amount of CSR achieved by WCPSS on growth in student achievement. The method tests the effect of CSR on the K-3 end-of-grade profile scores (posttest) while controlling for the score each student received the year before (pretest).

Remaining random differences in the pretest scores between CSR and comparison students were controlled by a statistical technique known as the analysis of covariance. Such a technique is necessary because if, by chance, the CSR students had lower pretest scores than the comparison students, the CSR students could also be expected to have lower posttest scores. The analysis of covariance tests whether the difference between the pre- and posttest scores (growth) is larger for the CSR students—and equally so across all levels of the pretest (growth pattern)—than for the comparison students after an arithmetic adjustment to equalize the pretest scores for each group.

Limitations of the Research

Not equal to the rigor of randomized experimental designs, the nonequivalent comparison group methods used in this evaluation should show effects where they are present if:

- Class sizes in the target schools and grades were reduced substantially below the average class size in the district. (Students in smaller classes are expected to have higher growth in achievement scores than those in classes that more nearly approximate the district average.)
- The procedures used to create the comparison groups successfully controlled individual and school-level factors that may influence outcomes.

Other national research on class-size reduction (projects STAR and SAGE) did not use the analysis of covariance. Although the methods that were used were appropriate in their respective contexts, **they may have missed the anomalous findings presented here because they did not test the interaction or homogeneous regression hypothesis that treatment effects are the same for all subjects regardless of pretest scores.**

Additional limitations of the research call for caution in the interpretation of results. The K-3 end-of-grade profiles used to assess achievement represent ratings assigned by teachers according to standardized procedures. They are not the results of independent,

objective testing. Independent and objective end-of-grade tests are not used until the third grade in WCPSS. Accordingly, some interaction between teacher expectations and the smaller class size might be expected: teachers might believe their students improved more than they actually did because the teachers knew they had smaller classes. Despite other limitations, independent and objective tests eliminate teacher perceptions as a factor accounting for observed results. However, because of the shell game of assigning teachers to classrooms (discussed earlier), a sizable proportion of the teachers and even the principals did not know where the smaller classes were.

The research discussed here was not intended to test whether smaller class sizes increase growth in achievement. That is already known to be true. The purpose of this research was to assess the extent to which improved growth in achievement occurred in WCPSS during the 1999-2000 school year as a result of the amount of CSR the district was able to achieve. Accordingly, it is especially important to focus on the amount of CSR that was achieved when reviewing the results on program effects presented later. **The primary evaluation issue is the amount of class size reduction that was achieved, not the amount of improvement in academic achievement that was observed.**

Measures of Academic Achievement

As just introduced, teachers use several independent measures to assess student achievement throughout the year. A student's final score on each measure is recorded in permanent district records at the end of each year. In this evaluation, a student's score in the spring before the first year of CSR (1999) was treated as a pretest, while the score at the end of the first year (2000) was treated as a posttest.

Between 1999 and 2000, WCPSS significantly modified the several rating scales used to assess achievement in **math** in order to agree with updates in the state curriculum. The use of non-identical pre and posttests can create significant difficulties for interpreting results. In addition, the **math** scales are not designed to measure cumulative academic growth throughout the lower grades. Instead they utilize a four-point scale *for each year* to assess whether a student is below grade level (values 1 and 2), at grade level (value 3), or above grade level (value 4). About two-thirds of the students in grades 1-3 received a rating of three on each of the new math measures, an increase of more than 20 percent over the ratings assigned in the previous year. **Consequently, the math measures are not suitable for evaluating academic growth in the lower grades and were not used in this evaluation.**

The three measures used to assess **language skill achievement** in WCPSS were analyzed in this evaluation. The first is the "reading-level" measure, which assesses cumulative growth throughout the lower grades (K-3). Each value of the 12 point rating scale is represented by numerous criteria that a student must fulfill in order to attain the corresponding rating score. Standard instructions guide teachers in assigning ratings. A student is expected to have attained a rating of between one and three by the end of the kindergarten year. By the end of the third grade a student is expected to have matured to a rating of between 10 and 12.

The second is the "book-level" measure, which represents a standardized test to determine the difficulty of the texts a student can read and comprehend. Standard instructions guide the administration and scoring of the test for both reading proficiency and comprehension. The scale metric is from 0 to 32, the values of which correspond to standard levels of difficulty of a text. The recorded values begin with zero (none yet) and are grouped subsequently from 1-2, to 3-4, and so forth. For this analysis the nodes were coded as 1.5, 2.5, and so forth. The specific coding is substantially arbitrary. In use by school personnel, there is no difference between reading at a book-level of one or a book-level of two, so the measure is actually a scale with 16 values and a zero starting point. Like the reading-level scale, the scale measures cumulative growth throughout grades 1-3.

The third is the "writing-level" measure, which is a 12-point rating scale that is structured and applied in the same manner as the "reading-level" measure. The beginning point of each measure represents a child who cannot yet read (or write). The end point of each scale represents a child who has become an independent reader (or writer), a level that all children are intended to reach by the end of the third grade. Table 4 depicts the approximate scale scores that are expected for each of the measures by the end of each of the lower grades.

Table 4: Expected final scores in each lower grade for measures of language skill.

Achievement Measure	Expected Range at End of Grade			
	K	1	2	3
Reading Level	1-3	4-6	7-9	10-12
Book Level	0-8	9-20	21-30	31-32
Writing Level	1-3	4-6	7-9	10-12

The scale values are intended to represent approximately equal skill distances spanning the gap between not yet reading (or writing) and independent reading (or writing) proficiency. Undoubtedly, the measures are not true interval measures in the sense that we cannot be certain that the distance between a score of one and a score of two is exactly equal to that between a score of 11 and a score of 12. However, the scales have underlying continuous variable interpretations. They have standardized instruction manuals for assigning ratings or scores, and perhaps more importantly they behave like interval measures. The correlation of pre- and posttests is strong. The overall correlation coefficients for pre- and posttests for all measures among second-grade students are shown in Table 5. A correlation of 1.00 would mean that the pretest score would perfectly predict the posttest score.

Table 5: Grade 2 correlation coefficients for language skill measures: 1999—2000.

Achievement Measure	Correlation Coefficient
Reading Level	0.776
Book Level	0.777
Writing Level	0.676

The weaker correlation between pre- and posttest scores for the "writing-level" measure is consistent with expectations. There is generally less reliability in rating writing samples than in rating reading performance. That there is a weaker correlation for the

writing scores also indicates the use of caution in interpreting the results for this measure. Results for the writing scale are not presented here.

Results

Grade Levels and Models Examined

CSR was implemented, inadvertently in some cases, in kindergarten through the third grade. No scores were available for kindergarten students in 1999 (pretest), and thus results for kindergarten are not presented. Class sizes were reduced in only one school for the third grade and thus no results are presented for the third grade. Finally, many fewer students received Models 2 and 3 than received Model 1. Accordingly, the analysis considers Model 1 in the first and second grade. Model 1 created a new class that was approximately equal in size to all the other classes of the target grade level in a school. This is comparable to the model that has received experimental testing. Table 6 summarizes the overall results for model 1 in grades 1 and 2.

Table 6: Overall results by grade level for model 1.

Model	Grade	Average Class Size CSR/CMP	Measure	Pretest Differences: CSR lower/higher	Is Growth Pattern Similar for CSR and Comparison♦	Significant Difference in Growth: CSR higher/lower	Effect Size
1	1	18.44/24	Reading	Lower	No	Higher	+1.6 months
			Book	None	No	Higher	+1.7 months
1	2	20/25	Reading	None	No	Higher	+1.2 months
			Book	None	Yes	None	NA

♦This column displays the results of the traditional test for interaction in the analysis of covariance or homogeneity of regression. When no interaction was found a "yes" was entered in the column to indicate that the growth pattern was similar across all pretest scores. A "no" indicates that significant interaction between CSR services and pretest scores was present.

Results for First-Grade Students—Reading Level

The average class size for those first-grade students in Model 1 was 18.4, while the average for first-grade classes among comparison students was about 24. Table 7 shows that the *improvement in growth for reading level scores was higher among CSR students for each level of the pretest score except two, where they are approximately equal.*

CSR students with a pretest reading level score of 1 grew 2.86 scale points to a reading level score of 3.86, while comparison students improved only 2.64 scale points to 3.64. The average posttest score and the average growth are higher for CSR students than for comparison students, despite lower average pretest scores on the reading measure for CSR students.

Table 7: Growth in reading level for first-grade students (Model 1).

Reading Level							
CSR Students				Comparison Students			
Number of Students	Pretest Score	Average Posttest Score	Average Growth	Number of Students	Pretest Score	Average Posttest Score	Average Growth
29	1	3.86	2.86	11	1	3.64	2.64
40	2	6.00	4.00	34	2	5.62	3.62
45	3	7.33	4.33	49	3	6.63	3.36
12	4	8.33	4.33	21	4	7.00	3.00
3	5	8.33	3.33	11	5	8.36	3.36
4	6	10.75	4.75	2	6	7.00	1.00
1	7	10.00	3.00	2	7	10.00	3.00
0	8	--	--	0	8	--	--
0	9	--	--	2	9	5.50	-4.50
1	10	11.00	1.00	0	10	--	--
N=135				N=132			
Average:	2.58	6.45	3.87		3.1	6.36	3.27
P = .002			ъ = 1.167		=		ъ = .627
P = NA			Intercept = 3.44				Intercept = 4.422

Effect Size

The magnitude of the difference between program and comparison groups is often referred to as the effect size. In this evaluation, effect size was calculated using the statistically adjusted posttest averages for the two groups rather than the raw averages reported in Table 7. The adjusted figures represent the averages after controlling for random differences in pretest scores that remained after the procedures used to select the comparison group. *First-grade students in Model 1 for the reading measures had a significantly lower pretest average than did the comparison students; thus controlling for these differences is especially important.*

For each group of students, growth is equal to the difference between the adjusted posttest average and the overall pretest average for all students in both groups. Accordingly, the CSR students grew 3.9168 points during the year (6.752 – 2.8352), while the comparison group grew only 3.3638 points (6.119 – 2.8352). Thus the CSR students grew 0.553 points more than did the comparison students. Dividing the difference by the total amount of growth for the comparison students during the year shows that the additional growth for the CSR students was equal to 0.1644 of one year of growth (.553 ÷3.3638 for those students in larger classes). *Based on a 10-month school year this proportion is equal to about 1.6 months. Finn and Achilles(1990) reported that growth in classroom averages for reading SAT test scores for first-grade students with*

reduced class sizes in Project STAR was equal to "at least 1 ½" additional months of instruction in larger classes.[1] The present results are strikingly similar.

Growth Patterns Differ Across Pretest Scores

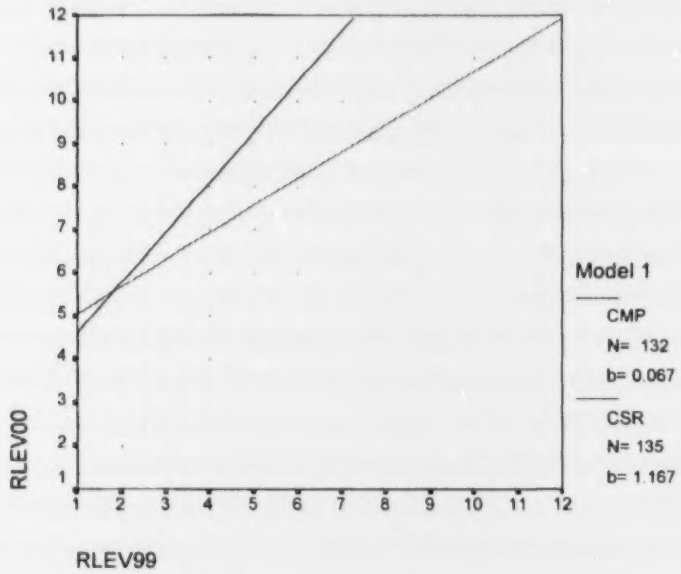
Improvement in academic growth for CSR students over comparison group students was greater for students with mid to high pretest scores than for those with low pretest scores. All students did not share equally the overall effect sizes reported in Tables 6 and 7. For example, students with a pretest score of one grew on average 0.22 points more than comparison students, while those with pretest scores of 4 grew on average 0.97 points more than comparison students. This pattern of results is referred to as ***interaction***. *It means that the pattern of growth across the range of pretest scores for CSR students was different, not just greater, than for comparison students.* According to strict adherence to statistical conventions it is not appropriate to report overall gains when differences in growth patterns are found. They are presented here to facilitate a direct comparison with the experimental research of Finn and Achilles. These researchers appropriately used posttest only measures because students were randomly assigned to small and normal classes. However, by using only posttest measures based on class rather than individual scores, this particular form of interaction may have been missed.

Figure 1 shows the data from Table 7 arranged as a scatter plot. The slope of the line representing CSR students (1.1667) is steeper than the slope representing comparison students (0.067). This illustrates that as pretest scores increased CSR students grew progressively more than did those in larger classes.

---

[1] Finn, Jeremy D. and C.M. Achilles. (1990). "Answers and Questions About Class Size." American Educational Research Journal. V 27, N 3, p. 567. The effect size can also be presented in terms of the standard deviation, which is how Finn and Achilles performed the calculation. The standard deviation is a measure of the amount of variation in scores within which about two-thirds of the students fall. A standard deviation of 1.42 indicates that about two-thirds of the students had scores that were no more than 1.42 scale points above or below the average. Accordingly, the adjusted posttest average for the comparison group (6.199) is subtracted from the adjusted posttest average for the CSR group (6.752). The difference (.553), which is equal to subtracting the raw growth scores reported in Table 7, is divided by the standard deviation of the overall average pretest score (1.4232) to show the difference in growth between groups in terms of pretest standard deviations (.3886 standard deviations). Accordingly, the CSR students grew .3886 more pretest standard deviations (0.553 ÷1.4232) than did the comparison group during the year. The comparison student growth (6.199 – 2.8352 = 3.3638) equals a total of 2.3635 pretest standard deviations during the year (3.3638 ÷ 1.4232 = 2.3635). Thus, the additional growth of the CSR students is equal to about .1644 of a year's growth for the comparison students (.3886 ÷2.3635) or, based on a 10-month school year, about 1.6 months of instruction in larger classes. The present calculation of effect sizes differs from that of Finn and Achilles who divided the difference of **posttest** means by the standard deviation of the comparison group **posttest** scores. Finn and Achilles converted the reported effect size based on posttest standard deviations into months of instruction by looking up the difference in "the [SAT] publishers table of norms, (p.567)" which related the observed effect size based on posttest standard deviations to the growth expected during one year. Since no such table existed for the measures analyzed here, and since pretest scores were specifically included in the analysis of covariance, the two steps were combined.

Figure 1: Growth pattern lines for first grade students (Model 1).



Results for First-Grade Students—Book level

The results for the book-level measure for the first grade were similar to those for the reading-level measure. The interaction was strongly present but the CSR students had higher growth at all pretest levels than did the comparison students.

Results for Second-Grade Students—Reading Level

The average class size achieved under Model 1 in the second grade was approximately 20, while the size of the comparison group classes were about 25. Table 8 compares growth for second grade students in CSR classes with comparison students. Similar to the students in the first grade, the average posttest scores and the average growth scores are higher for the students in CSR classes than for the comparison students. The data also show a similar pattern of interaction.

The results for the second grade show a less ideal result than those for the first grade. They show that while CSR students with mid to high pretest scores grew considerably more than the corresponding comparison students, those students with low pretest scores did not grow as much as comparison students. Overall however, CSR students grew more than did the comparison group students.

Results for Second-Grade Students—Book Level

The results for the book level measure did not follow the same pattern. For the book level measure, no interaction was present and there was no difference between the CSR students and the comparison students. These results may be due to the fact that the class sizes achieved were not as low in the second grade as they were in the first. More likely

they may result from ceiling effects in the book-level measure. Many second grade students already achieve the maximum scale score. Thus the growth of these children may well be beyond the capabilities of the measure.

Table 8: Growth in reading level for second-grade students (Model 1).

Reading Level							
CSR Students				Comparison Students			
Number of Students	Pretest Score	Average Posttest Score	Average Growth	Number of Students	Pretest Score	Average Posttest Score	Average Growth
7	1	3.57	2.57	8	1	4.25	3.25
15	2	4.13	2.13	15	2	4.80	2.80
26	3	5.58	2.58	20	3	5.85	2.85
66	4	6.99	2.99	66	4	6.74	2.74
76	5	7.72	2.72	122	5	7.62	2.62
123	6	8.63	2.63	129	6	8.74	1.74
151	7	10.06	3.06	93	7	9.69	2.69
65	8	10.46	2.46	51	8	9.96	1.96
63	9	11.24	2.24	66	9	10.76	1.76
50	10	11.54	1.54	57	10	10.95	0.95
20	11	11.60	0.60	29	11	11.17	0.17
24	12	11.92	-0.08	19	12	11.53	-1.53
N = 686				N = 673			
Average:	6.77	9.25	2.47		6.67	8.90	2.22
P = .006			ß = .791		=		ß = .696
P = NA			Intercept = 3.88				Intercept = 4.25

What Does Interaction Mean?

The observation that academic growth in smaller classes was less for students with low pretest scores than for those with mid to high pretest scores is unanticipated and provocative. What can it mean that students with very low pretest scores experience significantly less improvement in growth due to smaller classes than do students with mid to high pretest scores?

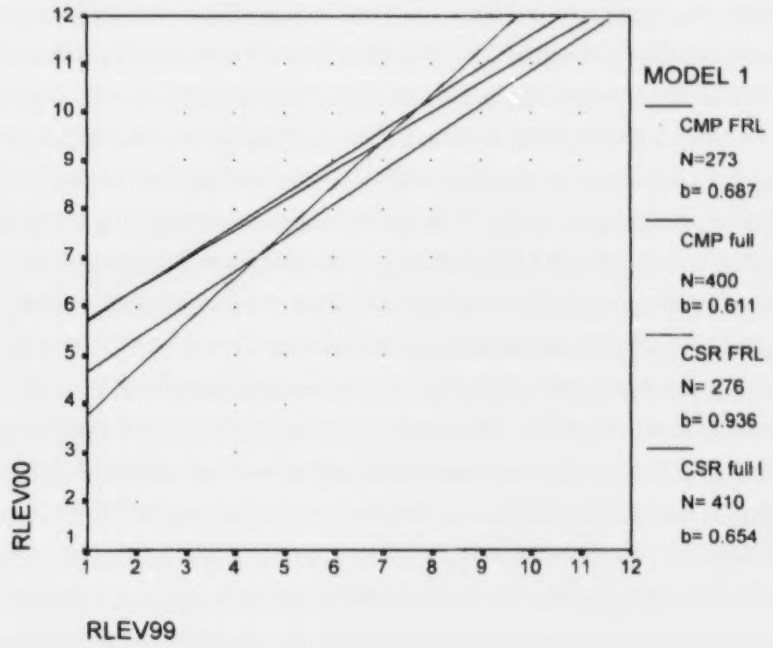
The author first consulted the research literature[2] on the many ways in which non-equivalent comparison groups can be contaminated, and how some patterns of results may produce equivocal interpretations. None of the confounding outcomes described in the literature appear to account for the observed result.

Exploration of the unanticipated results revealed that it was FRL students that accounted for the interaction. That is, in smaller classes, FRL students with low pretest scores did not grow as much more than their counterparts in larger classes, as did those FRL students with average or high pretest scores. Figure 2 displays the growth lines for each

[2] See for example Cook, Thomas D. and D. T. Campbell. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin Company.

group of students in model 1 for the second grade where the interaction was the most pronounced.

Figure 2: Growth pattern lines for second grade students (Model 1) by free or reduced- price lunch status.



The growth line that crosses all the others (showing interaction) represents CSR students with free or reduced-price lunches. It illustrates that among these students, those who began very low grew less than other categories of students, whereas those who began with mid to high range scores grew more than other categories of students. Overall, FRL students showed improvement over the comparison group equal to that of other students.

As discussed earlier, some of the comparison students were drawn from schools with a lower proportion of FRL students than the 23 target schools. As discussed, this occurred because there were not quite enough FRL students in the 23 target schools to complete the comparison groups. It seemed possible that these students might account for the observed interaction. To test this hypothesis the analysis was run separately for those students from schools with less than 30 percent FRL students and for those with more than 30 percent FRL students. The interaction was found in both groups. Moreover there was no overall correlation between growth scores for those students with low pretest scores (values = 1-3) and the proportion of FRL students in a school. That is, low-achieving students in general from schools with high proportions of FRL students did not appear to grow less than low achieving students from schools with lower proportions of FRL students.

Discussion

The federal class-size reduction program creates special implementation issues for districts and individual schools. Shifting enrollment can mean that the effect of class-size reduction moves from one grade to another as enrollment changes between hiring decisions and the arrival of students. Because the "squeaky wheel gets the grease" model is so entrenched, there is a tendency for class-size reduction to drift toward pullout programs for models intended for use with limited space.

The findings on academic achievement suggest that consistent improvements in language arts achievement can occur when class size is reduced from 24 or 25 down to 18-or 19 in the first grade. The findings were equivocal for second grade because class size was not reduced as much and because ceiling effects inherent in the measure used to assess reading comprehension may have obscured the actual program effects.

That free and reduced-price lunch students in smaller classes exhibit different growth patterns than other students is unanticipated and provocative. That FRL students with average to high pretest scores grow so much more than other students is good—this is what is needed to close persistent achievement gaps. That FRL students with low pretest scores do not grow as much as those with higher pretest scores and in some instances may not grow as much as similar students in larger classes is troubling. The possibility that this may occur calls for immediate steps to replicate this work and to determine why such an outcome occurs.

The research shows that better measures are needed to more adequately assess academic growth in the lower grades (k-3). A standardized test at the end of third grade is too late. Easy to apply measures used by teachers for both math and language arts need to represent continuous variables that show cumulative growth throughout the lower grades without a pronounced ceiling effect by the end of the second grade. Such data would empower teachers to adopt the data driven practices that are consistent with all continuous improvement and total quality management approaches.

In North Carolina, this would consist of developing math scales that are structured and applied in the same manner as the reading level scale. The book-level scale could be improved substantially by the addition of more difficult texts and corresponding scale points at the top end of the scale. All of the measures could easily be computerized to facilitate automated scoring of different dimensions of math and language arts achievement. This would provide teachers with a complete picture of each student relative to expectations at each assessment cycle during the year. I can envision a simple system that would identify each student's weak and strong areas and produce work assignments tailored for each student.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Evaluation of the Federal Class-Size Reduction Program in Wake County, North Carolina - 1999 - 2002

Author(s): David F. Scudder PhD

Corporate Source: Social Research & Assessment

Publication Date: April 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ↑ ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑ ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑ ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: D F Scudder	Printed Name/Position/Title: David F. Scudder, Ph.D.
Organization/Address: 1054 Chesterfield, Cir Winter Springs, FL 32708	Telephone: 407-963-5319 FAX:
	E-Mail Address: dscudder@cfl.rr.com Date: 4/4/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)